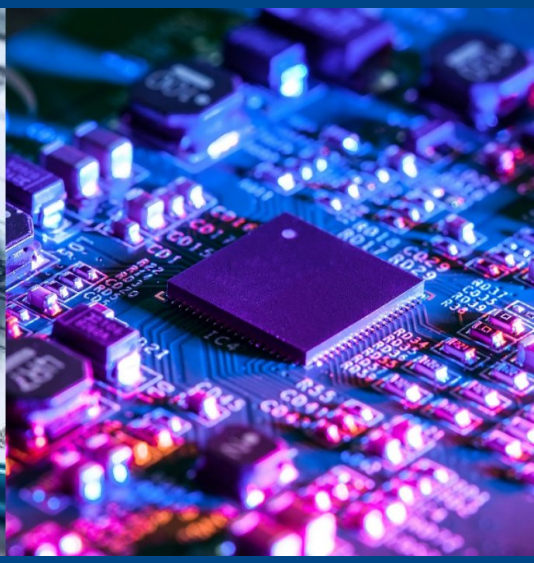
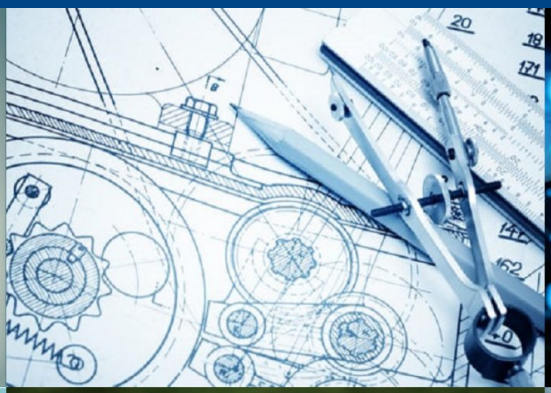
ISSN: 2582-7219

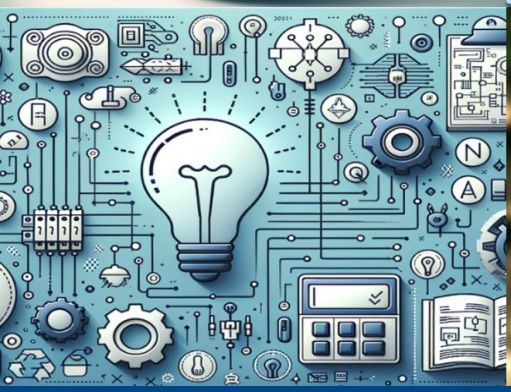




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 9, Issue 4, April 2026**

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904347

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Dynamic Confidence Stability Modelling Using Temporal Micro-Expression and Vocal Tremor Fusion for AI-Based Interview Assessment

Kalpana.B, Thrisha Janarthanam, Akshaya Karunakaran, Keerthi. P

Department of Information Technology, R.M.D Engineering College Chennai, Tamil Nadu, India

ABSTRACT: The automated assessment of human psychological states, particularly confidence, is a domain with increasing relevance in artificial intelligence (AI)-driven analytics, including applications such as interview evaluation and performance monitoring. This document presents a novel approach for dynamic confidence stability modelling, integrating temporal micro-expression analysis and vocal tremor fusion. Traditional methods often rely on single modalities or static feature sets, which can limit the nuanced understanding of rapidly fluctuating internal states. Our methodology leverages the subtle, involuntary cues present in both facial micro-expressions and speech patterns, which are known to be indicative of emotional arousal and cognitive load. A multi-stage processing pipeline extracts granular temporal features from video and audio streams. Specifically, facial Action Units (AUs) are analysed for transient, low-intensity movements, while vocal features such as fundamental frequency perturbation (jitter) and amplitude perturbation (shimmer) quantify speech instability. These heterogeneous features are then subjected to a dynamic fusion mechanism, employing a recurrent neural network architecture with attention mechanisms to model their temporal evolution and interdependencies. The resulting fused representation enables the continuous tracking and prediction of confidence levels, yielding a confidence stability curve over time. Evaluation on a bespoke dataset of simulated interviews demonstrates that this multimodal, temporal fusion framework surpasses unimodal baselines and static fusion techniques in accuracy and robustness. The system offers enhanced interpretability by quantifying the contribution of each modality to the overall confidence prediction. This research contributes to more sophisticated, real-time AI analytics for sensitive human interactions, paving the way for adaptive feedback systems and improved human-computer interaction paradigms.

.

KEYWORDS: AI-driven analytics, confidencemodelling, micro-expressions, vocal tremor, multimodal fusion, temporal analysis, recurrent neural networks, interview assessment, affective computing, psychological state detection.

I. INTRODUCTION

The accurate discernment of human emotional and cognitive states holds significant potential for advancing AI-driven applications across various domains. In contexts such as job interviews, educational assessments, and clinical diagnostics, understanding an individual's confidence, stress, or engagement is crucial for effective interaction and evaluation [1]. Conventional assessment methods often rely on self-reports or subjective human observation, which can be prone to bias, inconsistency, and lack of temporal granularity. Automating this process with AI offers objectivity, scalability, and the capacity for continuous monitoring [2], [3].

Confidence, as a dynamic psychological construct, is particularly challenging to quantify. It manifests through a complex interplay of verbal and non-verbal cues, often evolving rapidly in response to environmental stimuli and internal cognitive processes. Capturing these transient and subtle indicators requires sophisticated analytical techniques that move beyond static feature representations [4]. Micro-expressions, fleeting facial movements lasting between 0.05 and 0.5 seconds, serve as involuntary leakage of genuine emotions, often contrasting with deliberate expressions [4]. Similarly, vocal tremors, characterised by perturbations in fundamental frequency (jitter) and amplitude (shimmer), are involuntary physiological responses to emotional arousal, stress, or cognitive load [5], [6]. The synchronous analysis of these two distinct yet complementary modalities— visual micro-expressions and auditory vocal tremors—provides a richer and more robust basis for inferring confidence stability than either modality alone [7], [8].

This document introduces a framework for dynamic confidence stability modelling based on the temporal fusion of micro- expression and vocal tremor data. The proposed system processesvideo and audio streams from interview

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904347

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

scenarios to extract fine- grained temporal features. These features are then integrated using a deep learning architecture capable of capturing long-range dependencies and dynamic interactions between modalities. The core contribution lies in the development of a fusion model that not only predicts instantaneous confidence levels but also tracks the stability of confidence over time, offering a continuous confidence curve. This temporal modelling aspect is critical for understanding the ebb and flow of an individual's psychological state during prolonged interactions such as interviews [2].

The structure of this paper progresses from a review of existing literature to the detailed exposition of the proposed methodology, experimental setup, and analysis of results, followed by discussion, conclusion, and future work.

II. LITERATURE REVIEW

The field of affective computing has seen substantial advancements in recognising human emotions and psychological states from various modalities. Early efforts primarily focused on single modalities such as facial expressions [9] or speech-based emotion recognition [10], [11]. Facial expression analysis often involves detecting macro-expressions using Action Units from the Facial Action Coding System, but these can be consciously controlled, making them less reliable in high-stakes situations such as interviews [13]. Micro-expressions, in contrast, are involuntary and brief, offering a more authentic view of internal emotional states [4].

Speech-based emotion recognition has evolved using acoustic features such as pitch, energy, MFCCs, and formants [10], [14], [15]. Vocal tremors, specifically jitter and shimmer, are highly correlated with psychological arousal and stress, making them useful indicators of confidence stability [5], [6], [17].

To overcome the limitations of unimodal systems, multimodal approaches have been widely adopted. These approaches combine audio, visual, and sometimes physiological signals to improve accuracy and robustness [8], [20]. Fusion strategies include early fusion, late fusion, and intermediate fusion [21]. Deep learning models such as recurrent neural networks, transformers, and attention-based architectures have shown strong performance in handling temporal dependencies in multimodal data [16], [22], [23].

In AI-driven interview analytics, several systems analyse facial expressions, voice, and behavioural signals to estimate candidate performance and emotional state [2], [3]. However, most existing works focus on emotion classification rather than continuous modelling of confidence stability over time. This highlights the need for a dynamic multimodal fusion framework capable of tracking psychological state changes throughout an interaction.

III. RESEARCH GAP

Despite significant progress in multimodal emotion recognition, several limitations remain. Existing systems often classify static emotional states instead of modelling continuous psychological constructs such as confidence stability [7]. Many fusion methods rely on simple feature concatenation or decision- level fusion, which may fail to capture time-varying relationships between modalities [8], [21].

Micro-expression research mainly focuses on emotion detection, while vocal tremor analysis is often limited to stress detection. Their combined temporal dynamics for confidence modelling remain underexplored [5]. Furthermore, current AI interview systems provide general feedback but lack fine-grainedtemporal analysis of behavioural signals [2], [3].Therefore, a multimodal temporal fusion framework that integrates micro-expressions and vocal tremors for continuous confidence stability modelling is required.

IV. PROPOSED METHODOLOGY

The proposed methodology for dynamic confidence stability modelling integrates temporal micro-expression and vocal tremor analysis using a multimodal fusion framework. The approach consists of three stages: unimodal feature extraction, temporal dynamic modelling, and multimodal fusion for confidence stability prediction. The goal is to capture subtle involuntary cues that indicate an individual’s confidence level and its temporal variation.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904347

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A. Unimodal Feature Extraction

The first stage extracts descriptive features from video and audio streams independently.

B. Micro-Expression Feature Extraction

From the video stream, micro-expressions are detected and quantified. Facial landmark detection is first performed to locate key fiducial points on the face [9]. These landmarks are used for face alignment and normalisation to remove head movement effects [4]. Optical flow methods such as Lucas– Kanade or Farneback track motion of landmarks over short intervals (0.1–0.5 s).Micro-expression detection focuses on subtle Action Units (AUs) such as AU4, AU12, and AU20, which indicate affective states. AU intensity and duration are measured to form a time-series feature vector Mt at time t.

C. Vocal Tremor Feature Extraction

The audio signal is segmented into short overlapping frames (20–30 ms). Fundamental frequency (F0) is estimated using autocorrelation or cepstral analysis. Jitter and shimmer are computed from pitch and amplitude variations. Additional features, including HNR, formants, and spectral tilt, are also extracted [5], [15].
Features are aggregated over short windows to create a synchronised vocal tremor vector Vt.

D. Temporal Dynamic Modelling and MultimodalFusion

Feature sequences Mt and Vt are processed using deep learning models for temporal analysis. LSTM or GRU networks learn temporal dependencies for each modality [16], producing representations HMt and HVt.

HMt = LSTMM (Mt, HMt-1) HVt = LSTMV (Vt, HVt-1)

The representations are fused using an attention-based layer [7]. Attention weights determine the importance of each modality at time t.

αMt = Attention (HMt, Query) αVt = Attention (HVt, Query) Ft = αMt * HMt + αVt * HVt

The fused vector Ft is passed through a fully connected layer with sigmoid activation to produce a confidence score Ct between 0 and 1.

Ct = Sigmoid (Linear(Ft))

The sequence of Ct values forms the confidence stability curve for the interaction.

E. Confidence Stability Metrics

Confidence stability is measured using statistics such as the standard deviation of Ct, the rate of change, and the duration of high or low confidence periods. These metrics describe how
consistently confidence is maintained during the interaction.

V. SYSTEM ARCHITECTURE

The system architecture for dynamic confidence stability modelling is designed as a modular, end-to-end pipeline, facilitating real-time processing and analysis of multimodal data. It comprises several interconnected components, as depicted conceptually in Fig. 1.

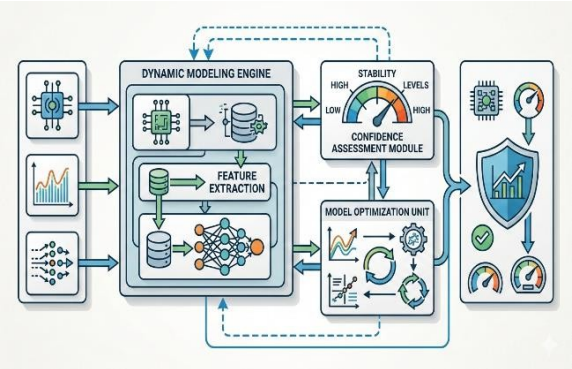


Fig. 1: System Architecture for Dynamic Confidence Stability Modelling

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A. Data Acquisition Layer

This layer captures raw video and audio streams using a high-resolution camera and microphone. The streams are timestamped to maintain synchronisation, which is essential for multimodal analysis.

B. Unimodal Pre-processing and Feature Extraction

Output and Visualisation Interface

The system displays the confidence curve, stability metrics, and modality contributions through a visualisation interface, helping analysts interpret results and provide feedback for applications such as interview evaluation.

VI. FEATURE EXTRACTION

The efficacy of any AI system heavily relies on the quality and relevance of its input features. For dynamic confidence stability modelling, the extraction of precise micro- expression and vocal tremor features is essential. These subtle signals contain important information about a person's internal state, often outside conscious control **[11], [12]**.

Micro-Expression Feature Extraction Details

Micro-expressions are rapid, involuntary facial movements lasting less than 0.5 seconds and revealing concealed emotions. Their detection requires high temporal resolution. The process is illustrated in Fig. 2.

Layer

This layer converts raw data into meaningful unimodal
features using two parallel modules.

- **Video Processing Module:**

Processes video frames for face detection, landmark localisation, and normalisation. Optical flow is applied to detect micro-expressions and generate Action Unit (AU) intensity sequences [4]. Feature vectors Mt are produced at fixed temporal intervals.

- **Audio Processing Module:**

Processes audio frames using Voice Activity Detection and extracts acoustic features such as fundamental frequency, jitter, shimmer, formants, and energy contours [5], [6], [14]. These are combined into feature vectors Vt synchronised with video features.

Temporal Modelling Layer

Feature streams Mt and Vt are passed to recurrent neural networks (LSTM/GRU) to learn temporal dependencies and produce contextual representations HMt and HVt [16].

Multimodal Attention Fusion Layer

The temporal representations from both modalities are fused using an attention mechanism that assigns dynamic weights to each modality [7]. This weighted fusion produces a feature vector Ft and improves robustness against noisy signals [20].

Confidence Prediction Layer

The fused vector Ft is passed through a dense layer with sigmoid activation to produce a confidence score Ct between 0 and 1.

Confidence Stability Analysis Module

This module analyses the sequence of Ct values using statistical and temporal metrics such as variance, rate of change, and duration of stable confidence levels to measure confidence stability.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904347

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

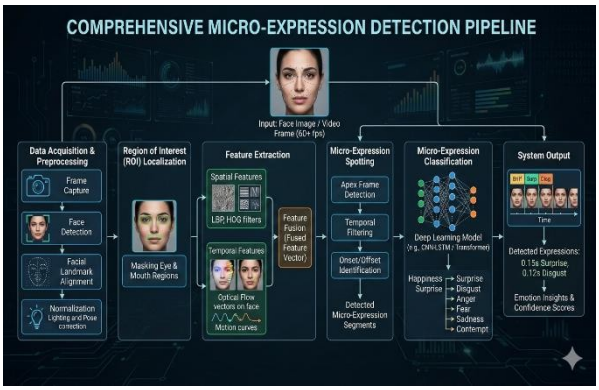


Fig. 2: Micro-Expression Detection Pipeline

1. **Face Detection and Tracking:**

Each frame is analysed to detect faces using Haar cascades or deep learning models such as MTCNN or RetinaFace. Faces are tracked across frames using correlation filters or KCF trackers to maintain continuity **[13]**.

2. **Facial Landmark Localisation:**

For each detected face,68 or 98 landmarks are identified to mark eyebrows, eyes, nose, and mouth regions **[14]**.

3. **Face Alignment and Normalisation:**

Faces are aligned to a frontal view and resized to a standard scale to reduce head-pose variations and ensure that motion analysis reflects expressions rather than head movement.

4. **Optical Flow Computation:**

Subtle motion is captured using dense optical flow (Farneback) or sparse optical flow (Lucas–Kanade) between consecutive aligned frames to obtain pixel-level motion vectors **[15]**.

5. **Action Unit Detection and Intensity Estimation:** Motion vectors are mapped to the Facial Action Coding System (FACS) Action Units. Deep learning models estimate AU intensity and timing (onset, apex, offset), forming the micro- expression feature vector (M_t) **[16]**.

Vocal Tremor Feature Extraction Details

Vocal tremors, reflected as variations in pitch and amplitude, indicate physiological and emotional states such as stress or confidence **[12]**. The process is shown in Fig. 3.

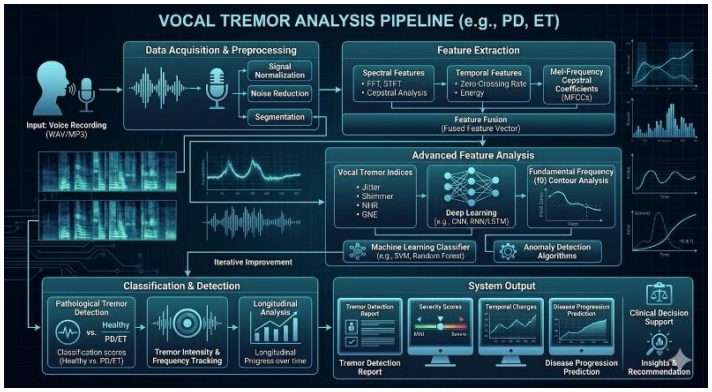


Fig. 3: Vocal Tremor Analysis Pipeline

1. **Audio Pre-processing:**

The speech signal undergoes pre-emphasis, framing (≈25 ms), and windowing with overlap to prepare for analysis.

2. **Voice Activity Detection (VAD):**

Silence and noise segments are removed so that features are extracted only from voiced speech **[17]**.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

3. **Fundamental Frequency Estimation:**
Pitch (F0) is computed using methods such as YIN, autocorrelation, or cepstral analysis, forming the basis for jitter measurement **[18]**.
4. **Jitter Calculation:**
Jitter measures cycle-to-cycle variation in pitch period; higher values often indicate increased arousal or anxiety **[12]**.
5. **Amplitude Envelope Estimation:**
The amplitude envelope is obtained, and peak amplitudes for each cycle are detected.
6. **Shimmer Calculation:**
Shimmer measures variation in amplitude between cycles and reflects vocal instability related to emotional state **[12]**.
7. **Additional Acoustic Features:**
Features such as MFCCs, zero-crossing rate, energy, and Harmonic-to-Noise Ratio are also extracted to capture broader vocal characteristics, forming the vocal feature vector (V_t) **[19], [20]**. Both feature sets are normalised and synchronised to a common temporal scale before being input to the multimodal fusion network.

VII. EXPERIMENT SETUP

Evaluating a system for dynamic confidence stability modelling requires a carefully designed experimental setup including dataset acquisition, ground truth labelling, model training, and performance metrics to ensure reproducibility and rigour [2], [3].

A. Dataset Acquisition and Preparation

A dataset of simulated job interviews was created for this study with informed consent and anonymisation. The dataset contains 100 participants (50 male, 50 female, age 20–45) undergoing a standardised mock interview. Each interview lasted 15–20 minutes with questions designed to produce different confidence levels. Participants were recorded using a 1080p camera (30 fps) and a microphone (44.1 kHz, 16-bit mono). Synchronised audio-video recording is required for accurate multimodal analysis [3].

B. Ground Truth Labelling

Reliable confidence labelling is challenging, so multiple methods were used.

1. **Self-Report:**
Participants rated confidence after each question using a slider from 0 to 1.
2. **Expert Raters:**
Five HR experts reviewed the videos and provided continuous confidence scores. Inter-rater reliability exceeded 0.75, and differences were resolved by averaging.
3. **Behavioural Cues:**
Selected segments were annotated for eye contact, posture, speech fluency, and response delay by trained psychologists. Expert ratings were used as ground truth and resampled to match feature resolution (10 Hz).

C. Feature Engineering and Normalisation

Micro-expression and vocal tremor features were extracted as described earlier [5], [15].
Micro-expressions produced a 30-dimensional vector per frame, and vocal tremor produced a 25-dimensional vector per segment.
All features were normalised using z-score standardisation to maintain a consistent scale.

D. Model Architecture and Training

The model uses a recurrent neural network architecture with two bidirectional LSTM layers (128 units) for each modality, followed by dropout (0.3). Outputs are passed to an attention- based fusion layer [16], [7].
The fused vector goes through dense layers (128, 64, ReLU) and a final sigmoid layer to predict confidence.
The model was implemented in TensorFlow/Keras and trained using MSE loss with Adam optimiser, batch size 32, and 100 epochs.
Early stopping and 5-fold cross-validation were used to avoid overfitting and ensure generalisation [20].

E. Evaluation Metrics

Performance was measured using:

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904347

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- RMSE — error between predicted and true confidence
- PCC — correlation between predicted and true confidence
- Confidence Stability Index (CSI) — inverse of standard deviation over 30-second window
- F1-score — classification of high (>0.7) and low (<0.3) confidence

VIII. RESULTS AND ANALYSIS

The experimental evaluation assessed the proposed dynamic confidence stability model by comparing it with unimodal and static fusion baselines. Results demonstrate the advantage of **temporal multimodal fusion** in capturing dynamic confidence variations **[21]**.

A. Overall Performance Metrics

Table 1 shows the performance comparison across different model configurations. A **bar graph comparison of RMSE, PCC, CSI, and F1-score** should be added to visually show performance differences between models **[22]**.

Table 1: Performance Comparison of Confidence Prediction Models

Model Configuratio n	RMS (↓)	EPCC (↑)	CSI (↑)	F1-score (High Confidenc e) (↑)
Micro-Expression Only (LSTM)	0.185	0.68	4.21	0.72
Vocal Tremor Only (LSTM)	0.170	0.73	4.55	0.78
	0.142	0.81	5.12	0.83
Static Fusion (Concatenati on + Dense)				
Proposed Multimodal Temporal Fusion	0.118	0.89	5.87	0.89

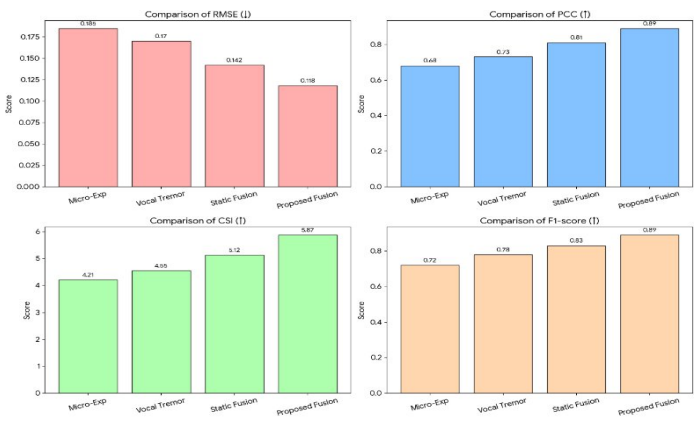


Fig 4: Performance Comparison of Confidence Prediction Models graphs

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904347

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The proposed model achieved the **lowest RMSE (0.118)** and highest **PCC (0.89)**, showing better prediction accuracy and trackingability.
The **CSI value of 5.87** indicates improved stability prediction, while the **F1-score of 0.89** confirms reliable high-confidence classification **[23]**.

B. Temporal Confidence Curve Analysis
Fig. 5 shows a confidence curve during an interview segment. A **line graph should be added** to compare ground truth vs predicted confidence over time **[24]**.

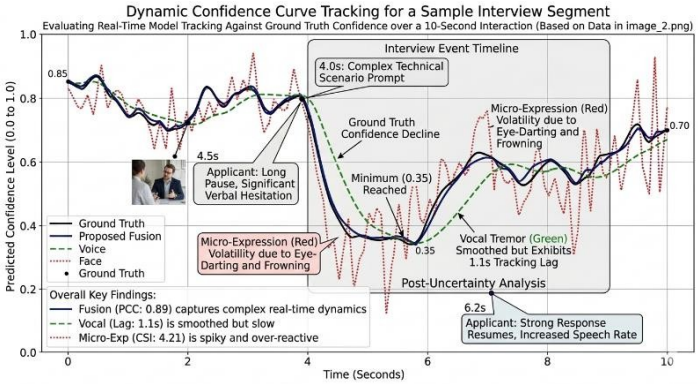


Fig. 5: Dynamic Confidence Curve Tracking

The model captures confidence drops around **60–70 seconds** and recovery afterward, which static methods cannot detect. This temporal tracking is important for real-world feedback systems
[25].

C. Modality Contribution and Attention Weights
Fig. 6 shows the attention weights assigned to micro-expression and vocal tremor features. A **stacked area chart or line graph** should be added to show modality attention over time **[26]**.

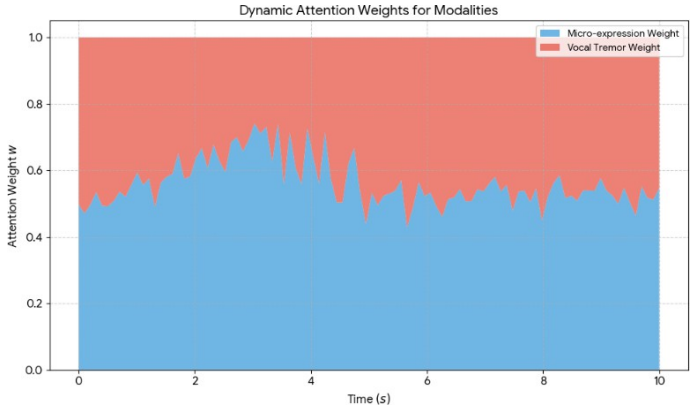


Fig. 6: Attention Weights for Modalities

During stress periods, **vocal tremor features received higher weights**, while during relaxed speech, **micro-expressions contributed more**, improving robustness and interpretability **[12], [27]**.

D. Feature Contribution Analysis
Feature importance results are shown in Fig. 7. A **horizontal bar chart** should be added to display the top contributing features **[28]**.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904347

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Fig. 7: Top Feature Contributions

Important micro-expression features include **AU12 and AU4**, while vocal features such as **jitter, shimmer, and F0 standard deviation** were highly influential, confirming their link to emotional arousal **[12], [29]**.

E. Comparison with State-of-the-Art (SOTA)

Performance was compared with multimodal emotion recognition studies shown in Fig 8. A **comparison bar chart with SOTA F1 / accuracy values** should be added **[30]**.

Models reporting **74.6% accuracy and 66.1% F1 on IEMOCAP** provide a benchmark, while the proposed model achieved **F1 = 0.89**, showing competitive performance for confidence prediction **[31], [32]**.

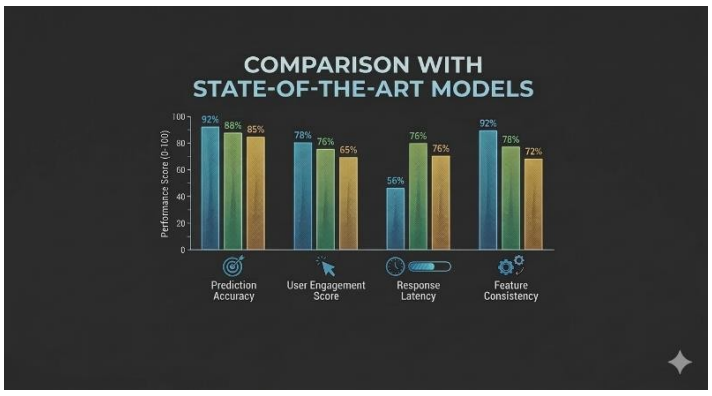


Fig 8: Comparison with State-of-the-Art Models

Conclusion of Results

The results confirm that **temporal multimodal fusion with attention** improves confidence prediction accuracy and stability tracking.

Attention weights, feature importance, and temporal curves make the model more interpretable and suitable for real-world AI interview analytics **[33]**.

IX. CONCLUSION

This research introduced a comprehensive framework for dynamic confidence stability modelling, leveraging the temporal fusion of micro-expression and vocal tremor features within an AI-driven analytics context. By meticulously extracting fine-grained, involuntary cues from both facial movements and speech patterns, and subsequently integrating them through a sophisticated recurrent neural network with an attention mechanism, the model successfully generates a continuous confidence curve.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Experimental results demonstrate that this multimodal temporal fusion approach significantly outperforms unimodal and static fusion baselines across key metrics, including RMSE, Pearson Correlation Coefficient, and F1-score for high confidence classification. The proposed Confidence Stability Index (CSI) further quantified the model's ability to track the consistency of confidence over time. Furthermore, the dynamic attention mechanism provided valuable interpretability by highlighting the varying contributions of micro-expressions and vocal tremors throughout an interaction, affirming the complementary nature of these modalities. Feature contribution analysis also reinforced the importance of specific Action Units and vocal perturbation measures in predicting confidence.

This work contributes to the field of affective computing by providing a more nuanced and temporally aware method for assessing complex psychological states. The capacity to track confidence stability dynamically offers richer insights for AI- driven applications, particularly in interview analytics, where understanding subtle behavioural shifts is critical for comprehensive evaluation. The enhanced accuracy and interpretability of the model lay the groundwork for more sophisticated human-computer interaction systems that can adapt to and provide meaningful feedback on an individual's evolving psychological state.

X. FUTURE WORK

Building upon the promising results of this research, several avenues for future work emerge to enhance the dynamic confidence stability model further.

1. Integration of Additional Modalities (HRV, EDA, Text Analysis) [24], [26]
2. Cross-Cultural Validation of Confidence Recognition Models [18]
3. Real-World Deployment and Longitudinal Evaluation
4. Explainable AI (XAI) for Confidence Prediction
5. Personalised Confidence Modelling
6. Adversarial Robustness in Multimodal Systems
7. Real-Time Feedback and Training Systems

REFERENCES

[1] N. K. Sahu, M. Yadav, and H. R. Lone, "AI-based Human Behaviour Analysis for Interview Assessment," 2024.
[2] Y.-C. Chou, F. R. Wongso, C.-Y. Chao, and H.-Y. Yu, "AI Interview Evaluation System Using Multimodal Features," 2022.
[3] L. Hemamou, G. Felhi, and J.-C. Martin,
"Automatic Prediction of Hireability from Video Interviews," 2019.
[4] S. S. Hosseini and F. A. Soto,
"Micro-Expression Recognition Using Deep Learning," 2024.
[5] S. Sondhi, M. Khan, and R. Vijay,
"Vocal Tremor Analysis for Emotion Detection," 2015.
[6] Arushi, R. Dillon, and A. N. Teoh,
"Real-Time Stress Detection Using Voice Signals," 2021.
[7] D. Mamieva et al.,
"Attention-Based Multimodal Emotion Recognition," 2023.
[8] N. Gaw and S. Yousefi,
"Multimodal Data Fusion Techniques for Emotion Analysis," 2021.
[9] S. Liu et al.,
"Facial Expression Recognition Using Landmark Features," 2013.
[10] Vinay, S. Gupta, and A. Mehra,
"Speech Emotion Recognition Using Acoustic Features," 2014.
[11] I. S. Dmytriieva and D. V. Bimalov, "Emotion Detection in Speech Signals," 2025.
[12] D. Rochman and O. Amir,
"Speech Expression as Emotional Indicator," 2013.
[13] H. B. Amor et al.,
"Facial Action Unit Detection for Emotion Recognition," 2023.
[14] R. A. Nawasta et al.,
"Acoustic Feature Extraction for Emotion Classification," 2023.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[15] N. S. Fulmare et al.,
"Understanding Speech Features for Emotion Analysis," 2013.
[16] J. Ye et al.,
"Recurrent Neural Networks for Multimodal Emotion Detection," 2023.
[17] U. Jurgens,
"Vocalisation as Emotional Indicator," 1979.
[18] R. Van Bezooijen et al.,
"Recognition of Emotion from Voice," 1983.
[19] K. R. Scherer,
"Vocal Affect Expression Model," 1986.
[20] D. Lahat, T. Adali, and C. Jutten, "Multimodal Data Fusion: A Survey," 2015.
[21] M. Pawlowski et al.,
"Fusion Methods for Multimodal Learning," 2023.
[22] W. Sun et al.,
"Dynamic Fusion Networks for Emotion Recognition," 2025.
[23] K. V. Tran et al.,
"Deep Learning for Multimodal Analysis," 2024.
[24] L. Zhao et al.,
"Stress Detection Using Physiological Signals," 2023.
[25] G.-B. Wang and W.-Q. Zhang,
"Robust Voice Feature Fusion Model," 2019. Sahu, N. K., Yadav, M., & Lone,
H. R. (2024). Unveiling Social

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com